Interconnection Networks for High-Performance Stream Computing with FPGA Clusters

氏名	アントネット　パンニリナン　モンディゴ Antoniette Pangilinan MONDIGO
学位の種類	博　士（情報科学）
学位記番号	情　博　第　702　号
学位授与年月日	令和　2年　3月25日
学位授与の要件	学位規則第4条第1項該当
研究科、専攻	東北大学大学院情報科学研究科（博士課程）情報基礎科学専攻
学位論文題目	Interconnection Networks for High-Performance Stream Computing with FPGA Clusters（FPGAクラスタによる高性能ストリーム計算のための相互接続網に関する研究）
論文審査委員	（主査）滝沢 寛之 教　授 張山 昌論 教　授　　佐野 健太郎 客員教授(RIKEN) 江川 隆輔 准教授

論文内容の要旨

High Performance Computing (HPC), as a field that relies heavily on cutting-edge technologies, is among the various domains affected by the impending end of Moore's Law and Dennard scaling. Many advancements in computing architecture across the various technology stack levels are being considered and explored to meet and support the growing demands of HPC applications. Field Programmable Gate Arrays (FPGAs), along with other dedicated acceleration platforms, are playing significant roles in this endeavor.

FPGAs, as reconfigurable devices, are recently seen as promising, energy-efficient hardware solutions due to a good balance between their flexibility and efficiency characteristics. Despite their typical lower operating frequency range than the other traditional platforms', creating custom hardware allows massively parallel operations with high utilization rates. Fine-grained and coarse-grained parallelism could be enabled and exploited by creating deep and wide computing pipelines with regular memory accesses. This allows continuous data streams to pass through the pipelines, while increasing the number of operations per memory access; therefore, fully utilizing the available bandwidth. All these make stream computing with a data flow model suitable for low operational intensity applications, such as stencil computing algorithms in FPGAs, which has been successfully demonstrated in numerous case studies. However, the resource budget of a single FPGA limits further performance scaling.

Just as clustered architectures dominate the current HPC trends, the utilization of FPGA clusters is a promising approach. However, despite numerous, successful case studies, FPGAs still lack widespread acceptance in general-purpose HPC installations. With this premise, it is the general direction of this research to aim at making FPGAs accelerators of HPC offloaded applications through system-wide custom computing. In order to achieve high performance, the main strategies are to increase the design space to support the increase of processing units, to reduce interconnection overhead, and to improve HPC application algorithms through customization. In this regard, stream computing with data flow is a suitable approach to fully exploit FPGA capabilities to meet high performance and high scalability demands in HPC.

Recently, large-scale deployments of FPGA clusters in data centers and cloud services have demonstrated the feasibility of providing a system-wide custom computing infrastructure

with FPGAs. However, the inter-FPGA interconnection network is an overhead-inducing region in the extended design space, which could affect the overall performance. Thus, there is a need to investigate a network's performance characteristics in order to achieve low-latency and high-throughput communication, especially for high-performance stream computing. Since the clustering architecture of FPGAs is typically selected based on target workloads and its required performance, choosing an appropriate interconnection network for FPGA clusters becomes an important aspect in this research. This dissertation is focused on investigating scalable interconnection networks for high-performance stream computing FPGA clusters. In particular, this dissertation focuses on the comparison of direct and indirect networks, with specific focus for stream computing requirements.

The main objective of this dissertation is to explore appropriate interconnection networks for high-performance stream computing FPGA clusters, where suitability and feasibility of direct and indirect networks are investigated. Direct networks are a common interconnect approach in existing FPGA clusters due to their low-latency and scalable characteristics. Indirect networks, on the other end, are not widely explored in FPGA clusters due to their overhead in communication latency, but promises a scalable and flexible connectivity in creating a custom network datapath, which is necessary for forward portability in HPC. In this dissertation, a 1D torus or ring topology and a tree topology with switches are adopted in investigating a direct network and an indirect network of FPGAs, respectively.

In Chapter 2, the requirements for stream computing in FPGA clusters are investigated. Most HPC applications, which include stream computing, require a scalable network architecture with a small FPGA footprint, and an efficient, low-latency, high-bandwidth communication. In addition, one functional requirement for stream computing is the support of backpressure signals in the backpressure-less channels of the inter-FPGA network. Another challenge is the synchronization of communicating FPGAs. This chapter proposes a lightweight and efficient hardware backpressure mechanism for direct and indirect inter-FPGA communication. This is done by creating a custom network protocol with credit-based flow control for backpressure propagation between communicating FPGAs. Furthermore, to achieve high-performance and highly-efficient communication, which is important to stream computing, it is the goal of this chapter to identify the proposed backpressure mechanism's design parameters and understand how they affect overall performance. While the hardware backpressure mechanism is implemented on a direct network in this chapter, the same design principles and mechanism apply for an indirect network, which are further discussed in Chapter 4.

Chapter 3 focuses on direct interconnection networks with high-speed transceiver links. Stream computing applications require low-latency and high-bandwidth communication. Since the hardware resource of a single FPGA is limited, one idea to scale the performance of FPGA-based HPC applications is to expand the design space with directly connected FPGAs. This chapter presents a scalable architecture of a deeply pipelined stream computing platform, where available parallelism and inter-FPGA performance characteristics are investigated to achieve a scaled performance. For a practical exploration of this vast design space, a performance model is presented and verified with the evaluation of a tsunami simulation application implemented on Intel Arria 10 FPGAs. Scalability analysis is also performed, where speedup is achieved when increasing the computing pipeline over multiple FPGAs while maintaining the problem size of computation. Performance is scaled with multiple FPGAs; however, performance degradation occurs with insufficient available bandwidth and large pipeline overhead brought by inadequate data stream size. An existing, hardware bandwidth-compression is applied to the communication links to mitigate the performance

degradation caused by the bottleneck-prone inter-FPGA links, which resulted to improved efficiency.

In Chapter 4, indirect networks with high-speed Ethernet switches are investigated. As FPGAs become a favorable choice in exploring new computing architectures for the post-Moore era, a flexible network architecture for scalable FPGA clusters becomes increasingly important in HPC. In this chapter, a scalable platform of indirectly-connected FPGAs is presented, where its Ethernet-switching network allows flexibly customized inter-FPGA connectivity. However, for certain applications such as in stream computing, it is necessary to establish a connection-oriented datapath with backpressure between FPGAs. Due to the lack of physical backpressure channel in the network, the Ethernet-switched network utilizes the custom credit-based network protocol with flow control introduced in Chapter 2 in order to provide receiver FPGA awareness and is tailored to minimize overall communication overhead, introduced by the variable latency in using Ethernet switches. To know its performance characteristics, necessary data transfer hardware on Intel Arria 10 FPGAs is implemented, and its communication performance is modeled, which is then compared to a direct network's. Results demonstrate that the connection-oriented Ethernet-switched network achieves equivalent performance to a point-to-point network for stream computing with large data sets, which suggests good performance and scalability for large HPC applications.

Through prototype implementations, obtaining performance characteristics, performance modeling, design space explorations, and performance evaluations, these different evaluation methods in this dissertation have demonstrated the suitability and feasibility of direct and indirect networks for stream computing FPGA clusters. Since stream computing applications generally process large data sets, streaming these sufficiently large data streams scale the performance linearly with more FPGAs on both direct and indirect network types, since they are able to achieve equivalent network throughput. Due to this, both direct and indirect networks would be good choices for inter-FPGA communication for high-performance stream computing. On the other hand, performance of insufficient data stream sizes on both network types demonstrates the communication latency as an overhead-inducing factor, causing degradation of performance. In this case, the indirect network's total transmission time is higher than a direct network's, in which latency dominates, therefore, negatively affecting the overall performance.

For future work, design space exploration should be done with the newly released Intel Stratix 10 FPGAs, where their transceiver links support 100 Gbps data rate. This implies an improved effective network bandwidth, which suggests better performance for both direct and indirect networks. Another area of future work is to provide a standard platform for FPGA cluster management, such as mapping of applications and network configurations. As a general direction, the indirect network provides a scalable and flexible infrastructure for high-level synthesis compilers and virtualization management of a large-scale FPGA cluster.

論文審査結果の要旨

近年，アプリケーションに応じて専用の回路を構築できる Field Programmable Gate Array (FPGA) の高性能計算分野における活用事例が多く報告されており，複数の FPGA を接続して大規模な FPGA クラスタを構成する研究も行われるようになってきた。アプリケーションに応じて FPGA クラスタを設計，構築するためには，その計算カーネルを効率よく実行する専回用路自体の設計に加えて FPGA 間の相互結合網の構成も検討する必要がある。しかし，直接通信網や間接通信網といった異なる種類の相互結合網に関して，その特性の違いが専用回路による計算に及ぼす響影はほとんど明らかにされていない。本論文では専用回路を用いた大規模並列ストリーム計算に着目し，それを効率よく実行するための FPGA クラスタの特に相互結合網の設計指針について論じたものであり，全編 5 章からなる。

第 1 章は緒論である。

第 2 章では，効率的なストリーム計算のために FPGA クラスタの相互結合網に求められる要件を検討している．受信側でのバッファあふれを回避する機能を有する FPGA クラスタ専用のフロー制御プロトコルを提案し，そのハードウェア設計，実装、および評価を行っている．その結果，性能への影響の大きい設計パラメータを見出し，その適切な値を探索することで性能とハードウェア量の間にトレードオフがあることを明らかにしている。そのような FPGA クラスタの相互結合網の設計空間探索はあまり報告されておらず，今後の高性能分野での FPGA の有効活用のための有用な成果である。

第 3 章では，一対一接続に基づく直接通信網で FPGA クラスタの相互結合網を構成し，その有用性を議論している。このために，深くパイプライン化されたストリーム計算 FPGA システムを提案し，その実装も行っている。性能モデルを定義して，実用的かつ効率的な設計空間探索を行っている。FPGA 間通信が性能ボトルネックになりやすいことから，通信データに可逆圧縮を適用し，見かけ上のバンド幅の向上を実現している。通信性能とストリーム計算の実行性能との強い相関を明らかにするとともに，FPGA クラスタの活用事例を示すものであり，今後の高性能計算分野での FPGA 活用の動機付けとなる重要な成果である。

第 4 章では，汎用の通信規格に基づいた Ethernet スイッチを用いる間接通信網で FPGA クラスタの相互結合網を構成し，その設計空間や性能特性を議論している。フロー制御機能を Ethernet プロトコル用に設計し，その実現に必要なハードウェア設計および実装も行っている。評価の結果から，達成可能な通信時間や実効バンド幅を明らかにし，ストリーム計算専用 FPGA クラスタを構築した場合の性能推定を行っている。その結果として，十分に大きなデータを扱うストリーム計算の場合には，Ethernet の利用による間接通信網でも直接通信網と同等の実行性能を実現できることを示しており，FPGA クラスタの構築に既存の汎用な通信技術を利用できることを示す実用性の高い成果である。

第 5 章は，本論文を総括し，結論としている。

以上要するに本論文は，FPGA クラスタにおける相互結合網に求められる性能要件の明確化，設計空間探索，および実装のすべてを行い，そのストリーム計算に対する有用性を論じてまとめたものであり，情報基礎科学および計算機科学の発展に寄与するところが少なくない。

よって，本論文は博士（情報科学）の学位論文として合格と認める。